View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repositorio Institucional Universidad de Medellín

A multi level approach for business process retrieval

Cristhian Figueroa[*]
David Camilo Corrales[**]
Juan Carlos Corrales[***]

Recibido: 10/03/2014 • Aceptado: 12/12/2014

Abstract

Nowadays business process reuse is critical in companies that need to build flexible and service-based business solutions in order to react quickly and cost-effective to dynamic market-conditions. For this reason, many companies have implemented approaches to find relevant business processes to be reused to create new software solutions performing required business functionalities. This paper presents a multilevel retrieval approach that detects linguistic, structural, and behavioral properties to increase the precision level in recovering those business processes stored in a repository.

Key words: business process, behavioral semantics, sub-graph isomorphism, control-flow patterns, relevance analysis.

[*] M.Sc. in Telematics Engineering, and Researcher of Software Engineering Group at Politecnico di Torino, Italy. Address. Corso Duca Degli Abruzzi 24, 10142, Turin, Italy, Tel. + 39 011 090 Ext. 7087. E-mail. cristhian.figueroa@polito.it.

[**] M.Sc. in Telematics Engineering, and Researcher of Telematics Engineering Group and Environmental Study Group at University of Cauca, Colombia. Address. Carrera 2 N.° 1A-25. Popayán, Colombia. Tel. +57 (8) 209800 Ext. 2145. E-mail. dcorrales@unicauca.edu.co

[***] Doctor of Philosophy in Sciences, Specialty Computer Science, and Full Professor and Leader of the Telematics Engineering Group at University of Cauca, Colombia. Address. Calle 5 N.° 4 - 70. Popayán, Colombia. Tel. +57 (8) 209800 Ext. 2129. E-mail. jcorral@unicauca.edu.co

Un enfoque multinivel para la recuperación de procesos de negocio

Resumen

Actualmente reutilizar procesos de negocio es un procedimiento crítico especialmente para compañías que requieren construir soluciones flexibles y soportadas por servicios con el fin de afrontar de manera efectiva y a bajo costo las condiciones cambiantes del mercado. Por esta razón, muchas de ellas han implementado métodos para encontrar procesos de negocio relevantes que puedan ser reutilizados en la creación de nuevas soluciones software que cumplan con una determinada función de negocio. Este artículo presenta un método multinivel que detecta similitudes entre procesos de negocio, teniendo en cuenta propiedades lingüísticas, estructurales y de comportamiento, con el fin de incrementar el nivel de precisión en la recuperación de aquellos procesos existentes en un repositorio.

Palabras clave: procesos de negocio, semántica del comportamiento, isomorfismo de sub-grafos, patrones de flujo de control, análisis de relevancia.

INTRODUCTION

The large amount of existing Business Processes (BP) in enterprise repositories has generated the need for effective retrieval mechanisms to find BP in order to build a new one by reusing it, or to be executed when required. BP retrieval is one of those technologies, which is a collection of techniques for locating BP based on different BP properties. This paper presents BeMantics (Behavioral Semantics BP retrieval) a multilevel retrieval approach that detects linguistic, structural, and behavioral properties of the BP to increase the precision level of the matching between a query BP and a set of existing BP [1]. Additionally this paper evaluates the relevance of BeMantics by comparing its results with other BP retrieval tool called StCoBP [2]. The relevance is evaluated through the precision, recall, averaged normalized discounted cumulated gain (ANDCG), and generalized average precision (GenAveP') measures from the information retrieval (IR) field.

On section 2, an overview of the current approaches to BP retrieval is presented. BeMantics, our multi-level approach to BP retrieval with enhanced matching precision based on behavioral semantics analysis is presented on section 3. The evaluation process of the proposed approach is described in section 4, while its results are provided in section 5. Finally the conclusions of the authors are given.

1. CURRENT APPROACHES FOR BP RETRIEVAL

Nowadays, BP retrieval techniques can be classified according to the BP properties in four levels: interfaces, semantics, structure and behavior. The first one takes into account parameters related with inputs/outputs and names of each task of BP [3, 4]; the second one is focused on the semantic inference of related concepts contained in ontologies [5-8]; the third one, compares the BP structure usually represented trough modeling formalisms which ease the structural analysis using mathematical techniques, such as graph isomorphism [9-12]; and the last one, compares the behavior of BP represented as interchange of messages within tasks, record of historical execution of BP, and control-flow [13, 14].

However it has been demonstrated that in many situations, the application of those levels separately it is not enough to achieve relevant results that satisfy users' demands, thus, it's required to use the sum of the contributions of each single level [15].

One multi-level approach that combines two levels is named StCoBP (structural comparison for BP), developed by Corrales, Gomez and Corrales [2], detects equivalences between a query BP and a set of BP from a repository using structural and linguistic properties. It receives a query BP described by using the BPEL4WS language

(BP execution language for semantic WS) which enriches the interfaces (inputs/outputs) of the BP. Subsequently, the query BP is transformed to a formal representation based on graphs (BP graphs). In this way, the StCoBP compares one BP graph with a set of BP stored in a repository, known as network of BP [16] in order to find equivalences between them. To do that StCoBP is based VF2 algorithm which executes a sub-isomorphism comparison and is used to estimate the structural similarity according to a set of correspondence categories, which are applied when a query BP is compared with each one of the BP from repository. Therefore, if the BP query and a BP from the repository are structurally identical then the category is exact; if the BP query is contained in the BP from the repository then the category is plug-in; if the BP query contains the BP from the repository then the category is subsume; and otherwise the category is fail.

StCoBP assigns different values of structural similarity according to the categories described above; however, this approach does not take into account the behavioral level which is important in order to discover BP according to the way in which it can be executed.

2. BEMANTICS, A MULTILEVEL APPROACH FOR BP RETRIEVAL

BeMantics is an approach for retrieving BPs that not only detects equivalences between BPs using structural and linguistics properties, but also the behavioral properties. The main difference with respect to StCoBP is a pre-matching phase used by BeMantics in order to filter BP (i.e. to generate a pre-ranking of BP) according to the behavioral semantics property which has into account the detected control-flow patterns and their semantic relations depicted in a control-flow patterns ontology.

The BeMantics approach receives semantically annotated BP described in the BPMO (business process modeling ontology) language which is based on the BPMN notation (BP modeling notation). Therefore, the semantics comparison estimates a linguistic (lexical and semantics) distance between names and interfaces of two BP. Regarding, the structural comparison, it has a slightly difference with the StCoBP one, because d their r uses the A* algorithm to find inexact matching by editing nodes and edges of the BP in order to generate a BP similar to a query BP; while StCoBP uses the VF2 graph isomorphism algorithm which only find exact matches between substructures of the compared BP. Additionally, the BeMantics approach includes a behavioral semantics BP repository which executes an indexing mechanism based on control-flow patterns and its semantic relations. The main modules of this approach are presented in figure 1.

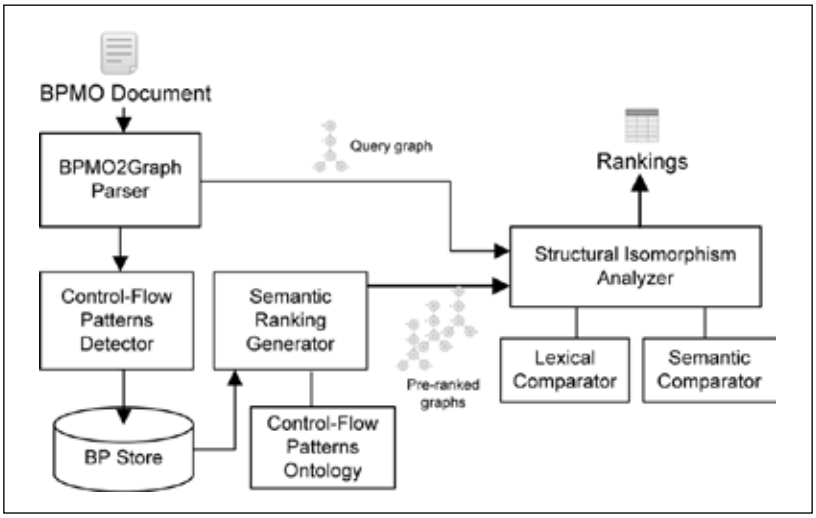

Figure 1. Architecture of the BeMantics approach.

Source: authors

- **BPMO2Graph parser:** it transforms a BPMO process to java objects using the WSMO4J API[1], and then applies transformation rules to get a process graph composed of activities nodes, connector nodes (AND, OR, and XOR) and edges to link them.
- **Control-flow patterns detector:** this module receives as input a process graph and returns a set of detected patterns. The control-flow patterns are sub-structures composed of activity nodes, connector nodes and edges representing a specific execution behavior of the BP. For example, figure 2 shows an example of the detection of two patterns, the e*xclusive choice* and the *sequence pattern* in a BP. As can be seen the control-flow patterns are only sub-structures that can be detected inside the full structure of the BPs

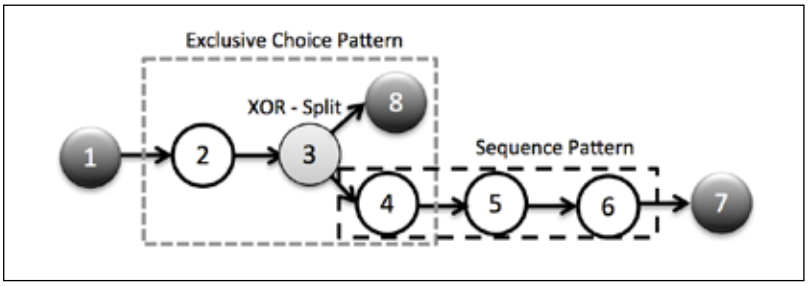

Figure 2. Example of detection of the patterns sequence and exclusive choice.

Source: authors.

[1] http://wsmo4j.sourceforge.net

Accordingly, the patterns detection is addressed by an isomorphism algorithm which finds sub-graphs structures (sub-structures representing the control-flow patterns) within a graph (the full structure of the BP) [16].

- **BP Store:** this module stores the BP and includes an index mechanism based on the set of patterns detected in each stored BP.

- Control-flow patterns ontology: the control-flow patterns ontology is a set of concepts representing control-flow patterns and their relations.

- Semantic ranking generator: this module uses the control-flow ontology in order to calculate a semantic distance between the control-flow patterns and consequently between the BP which contains a similar set of patterns. The semantic similarity and the number of detected patterns in each BP are later used to generate a ranking of BP. This module receives a set of detected control-flow patterns (PQ) in a query BP, and returns a ranking list of the stored BP in the repository, which have a similar set of control-flow patterns PT.

Additionally, this filter uses a control-flow patterns ontology in order to increase the search space by finding other BP having a patterns with similar functionality but not necessary the same. We called this as semantic analysis of control-flow. It means, that this module performs an inexact behavior detection based in the control-flow patterns ontological relationships. The behavioral semantics similarity used to evaluate the similarity between two BPs according to their similar patterns can be measured using a similarity function: $Sim_{Pattern} = \frac{\left|P_Q \cap P_T\right|}{\left|P_Q\right|}$.

- Structural isomorphism analyzer: this module receives as input a query graph and the set of pre-ranked BP obtained by the semantic ranking generator. The pre-ranked BP set are then structurally matched with the query graph using a graph isomorphism mechanism based on the A* algorithm.

- Lexical comparator: it is responsible for comparing the labels of the nodes names using the lexical mechanisms as Ngram, check synonym and check abbreviation [17].

- Semantic comparator: it compares the labels and interfaces (input/outputs) using semantic inference over the concepts from domain ontologies. This comparison is estimated by calculating the semantic distance between nodes semantically annotated with concepts of the domain ontologies.

3. EVALUATION

3.1 Reference benchmarking

The reference benchmarking tests the relevance measures for the StCoBP and the BeMantics approaches. A BP test set was created with 60 BP from the telecommunications domain and 40 BP from the geo-processing domain modeled using the BPMO and BPEL4SWS languages. Additionally a pertinence evaluation model [18] was designed in order to ease human judges to issue relevance judgments for BP obtaining a set of relevant BP. The set of relevant BP is used later in order to catalogue retrieval approaches by their level of retrieval effectiveness. Effectiveness is given if BP tools and evaluators come up with a similar result.

3.2 Relevance measures

The relevance of the results obtained by BeMantics and StCoBP are evaluated using the recall (Rg) and precision (Pg) measures. Precision (equation 1) estimates the ability of the system to retrieve only relevant elements (i.e. elements considered as relevant by human judges), while the recall (equation 2) evaluates the ability of the system to retrieve all the relevant elements avoiding missing relevant results.

$$P_g = \frac{\sum_{T_i \in T} \min\{f_r, f_e\}}{\sum_{T_i \in T} f_e} \quad R_g = \frac{\sum_{T_i \in T} \min\{f_r, f_e\}}{\sum_{T_i \in T} f_r} \quad (2)$$

Equations 1 and 2, shows the precision and relevance measures where T_i represents each element of the stored BP; f_e is a ranking generated by the BeMantics or StCoBP approaches; and the f_r is the set of relevant BP obtained by the human judges evaluation. The precision and recall are based on the number of relevant BP, however those measures does not take into account the position of the results in the ranking; for this reason this paper studies the ANDCG (average normalized discounted cumulated gain) and GenAveP (generalized average precision) measures presented and improved by Küster and König-Ries [19] which quantify the quality of the rankings produced by the BP retrieval tools based on the cumulated gain.

Equations 3 and 4, presents the *ANDCG* and *GenAveP*' measures where *CG* is the cumulated gain, $ICG(i)$ is the ideal CG, DCG is the discounted CG, $IDCG$ is the ideal DCG. The $CG(i) = \sum_{j=1}^{i} g(r_j)$, cumulated gain at rank i, measures the gain (g) that an automatic tool assigns to the top i items in a ranked list. The $ICG(i)$ measures the gain that a user assigns to the top i items in a ranked list (relevant BP).

The $DCG(i)=\sum_{j=1}^{i}\frac{g(i)}{disc(i)}$ is similar to the CG but uses a discount factor ($disc(i)$) which gives more value to the first elements and reduces the value to the latest elements in a ranked list, in this paper the discounted factor used is: $disc(i)=\max(1,\log_b i)$.

$$ANDCG=\frac{1}{|R|}\sum_{i=1}^{|L|}\frac{DCG(i)}{IDCG(i)} \quad (3)$$

$$GenAveP'=\frac{\sum_{i=1}^{|L|}\frac{CG(i)}{i}}{\sum_{i=1}^{|R|}\frac{ICG(i)}{i}} \quad (4)$$

As can be seen in equations 3 and 4, the *GenAveP*' is similar to the $ANDCG$ equation, but in the first one the discounted factor is the position i of an item in the ranking; the CG is evaluated for the set L of the first i items returned in response to a query, and the ICG for the set R of relevant items regarding a query.

3.3 Evaluation criteria

The evaluation criteria provide numerical values to facilitate human judges to issue similarity judgments between BP from a test repository. These relevance judgments are considered a hierarchical set of reliable relevant BP [18]. In this paper, the evaluation criteria are classified in two categories: structure and linguistic. The structural criterion analyzes the graphic structure and dependence relationships between activities of a BP; while the linguistic criterion takes into account the semantics and lexical features of the tasks names, descriptions and interfaces (input/output).

4. RESULTS

This section presents the main results by comparing the performance and the relevance for the BeMantics and StCoBP approaches. Below those analyses are described in detail.

4.1 Performance Analysis

The performance analysis estimates the average time execution of the StCoBP and BeMantics approaches (figure 3).

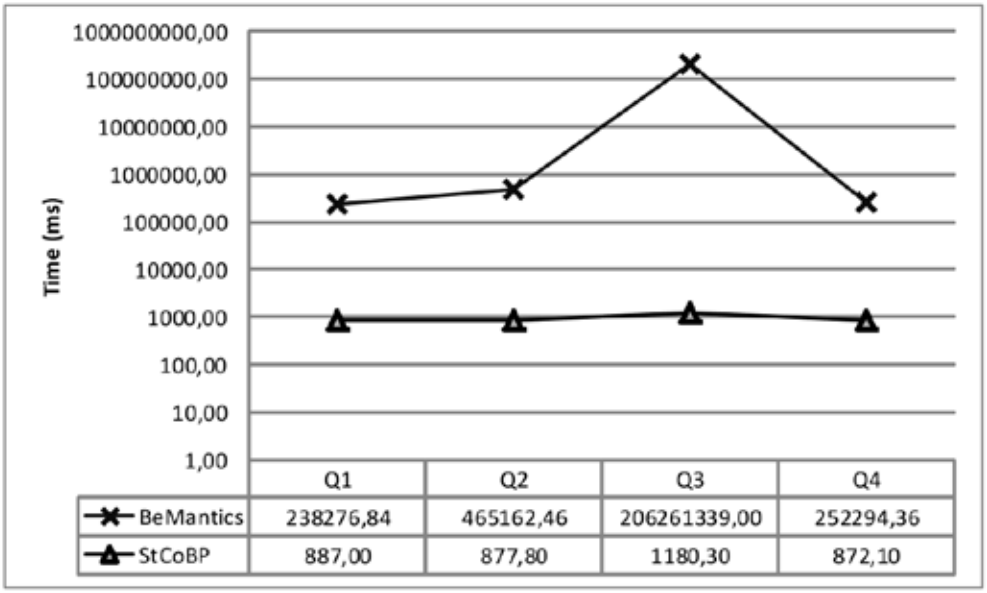

Figure 3. BeMantics vs StCoBP: Time execution for each query (Q1, Q2, Q3, and Q4)

Source: authors

Figure 3 presents, in a logarithmic scale, the evaluation of the two approaches using four queries BP (Q1, Q2, Q3, and Q4) . The performance results shown the StCoBP approach was more efficient than BeMantics approach, because BeMantics uses the A* algorithm, which consumes more execution time when the numbers of nodes to be compared is high [12]. On the other hand, StCoBP uses the VF2 algorithm which presented favorable results for sparse graphs, because its functionality is based on deterministic correspondence [20], obtaining a better performance than A* algorithm.

4.2 Relevance Analysis

- Precision and recall: In this paper the precision and recall measures are used in order to estimate the ability of the retrieval tools to find relevant BP from a set of BP stored in a repository. Figure 4 shows that BeMantics scored the highest Pg values for the structural (0.74) and the linguistic (0.87) criteria, and StCoBP the lowest values (0.35 and 0.36). This is because StCoBP used a BPEL4WS version of the original BPMO processes developed originally in the BeMantics approach, and in this case when the BPMO processes were transformed to BPEL4WS some inputs/outputs where missed, producing a reduction in the Pg for the linguistic criterion which has into account all the inputs/outputs names. Additionally, in the

structural criterion the StCoBP approach scored lowest values because it modified the VF2 algorithm by not only detecting the query BP in the network of BP, but also detecting substructures of network BP in the query BP. For this reason, the StCoBP approach retrieves more BP (relevant and no relevant) having a greater Rg values (0.35).

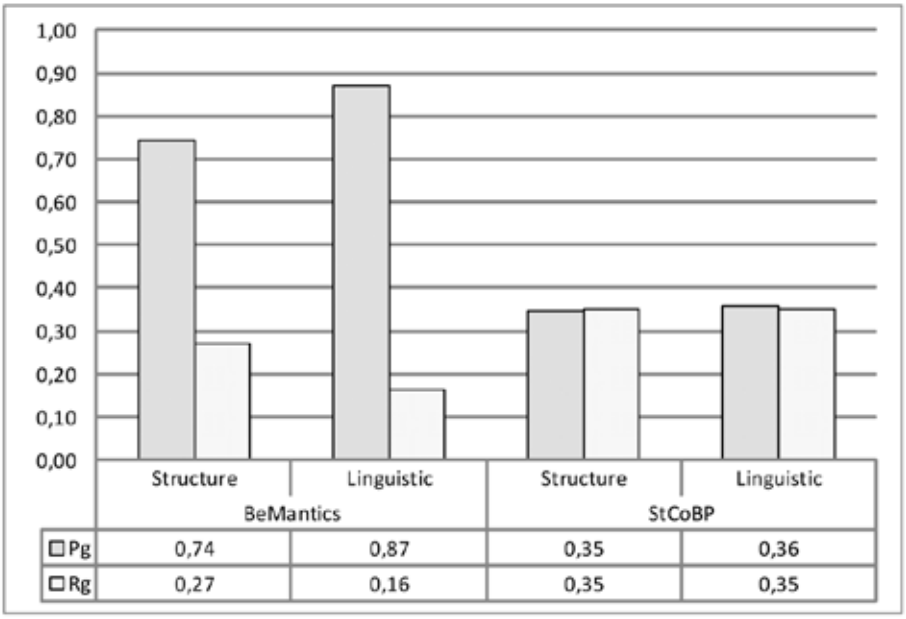

Figure 4. BeMantics vs StCoBP: Precision and recall.

Source: authors

- ANDCG and GenAveP': Regarding the ANDCG and GenAveP measures, figure 5 shows that StCOBP has better ranking quality than BeMantics. This is because unlike BeMantics, the StCoBP approach assigned greater values for the first elements and reduces the value to the latest elements in the output ranked list. Therefore, it can be observed that although BeMantics was more accurate (greater values of Pg) than StCoBP, this last one arranged the ranking results in a similar way than human judges did. This is important having into account persons put more attention on the first items of a search results.

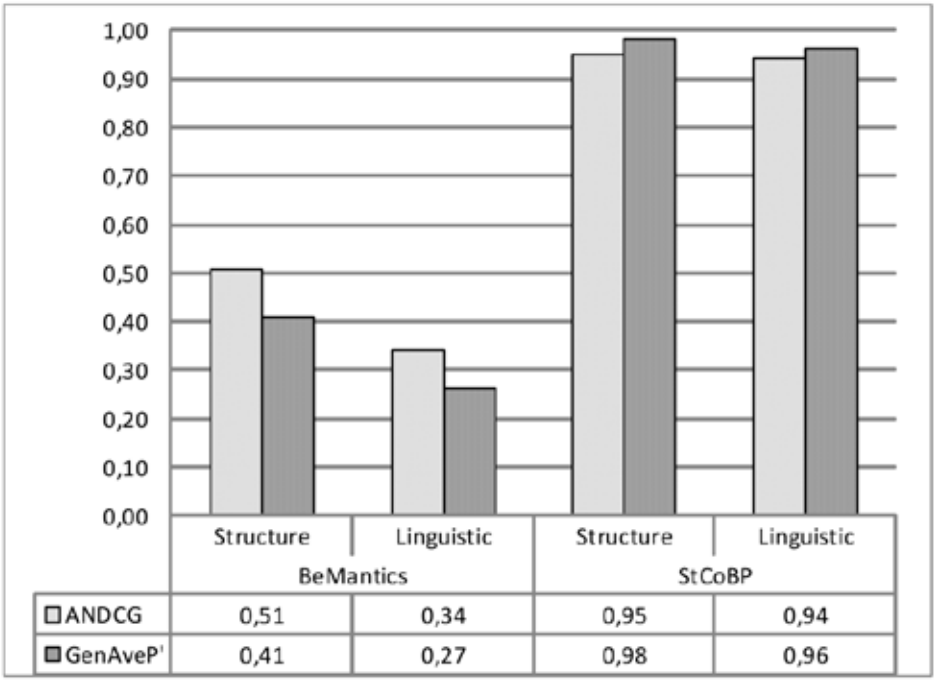

Figure 5. BeMantics vs StCoBP: ANDCG and GenAveP

Source: authors

CONCLUSIONS

This paper describes an approach for BP retrieval called "BeMantics". It's based on finding and analyzing behavioral properties of the BPs in order to increase the precision of the results. The relevance of the results was evaluated using measures from the information retrieval (IR) domain such as precision, recall, GenAveP and ANDCG. The evaluation of relevance and performance showed that BeMantics scored a high precision value (0.78, 0.87) but with high memory consumption cost and execution time; and StCoBP presents a less execution time but lower precision values (0.35, 0.36). Hence, the need for new intelligent approaches arises in order to obtain many relevant results avoiding irrelevant ones in acceptable time consumption. On the other hand, many approaches for discovering business processes, web services and information, have evaluated their results based on recall and precision measures, however, as demonstrated in this paper these measures estimate only the ability of the retrieval approaches to find relevant results, but regard-less of the position in which they appear in the output ranking. Therefore using measures such as GenAveP and ANDCG can be obtained more information about the quality of rankings of the retrieval tools.

ACKNOWLEDGMENTS

The authors would like to thank COLCIENCIAS for supporting the research projects that were the base for this paper. The authors also would like to thanks to the Universidad del Cauca and Telematics Engineering Group for supporting this research.

REFERENCES

[1] C. Figueroa and J. C. Corrales, *Recuperación Multinivel de Procesos de Negocio basada en Semántica del Comportamiento*. New York: Research and Innovation, 2012.

[2] D. C. Corrales, J. E. Gomez, and J. C. Corrales, *Comparación estructural y lingüística de procesos de negocio semánticos*. New York: Research and Innovation, 2012.

[3] A. Koschmider, T. Hornung, and A. Oberweis, "Recommendation-based editor for business process modeling", *Data & Knowledge Engineering,* pp. 483-503, June 2011.

[4] E. Goncalves, L. Ferreira, and M. Sinderen, "Towards runtime discovery, selection and composition of semantic services", *Computer Communications,* pp. 159-168, 2011.

[5] P. Châtel, "Toward a semantic web service discovery and dynamic orchestration based on the formals specification of functional domain knowledge", *20th International Conference on Software & Systems Engineering and their Applications (ICSSEA'2007),* 2007.

[6] L. Lin and B. Arpinar, "Discovery of semantic relations between web services", *ICWS '06 Proceedings of the IEEE International Conference on Web Services,* pp. 357-364, 2006.

[7] C. Kiefer, A. Bernstein, H. J. Lee, M. Klein, and M. Stocker, "Semantic process retrieval with isparql", *Proceedings of the 4th European conference on The Semantic Web: Research and Applications,* pp. 609-623, 2007.

[8] M. Klusch and F. Kaufer, "Wsmo-mx: A hybrid semantic web service matchmaker", *Web Intelligence and Agent Systems* pp. 23-42, January 2009.

[9] D. Grigori, J. C. Corrales, M. Bouzeghoub, and A. Gater, "Ranking BPEL processes for service discovery", *IEEE Transactions on Services Computing,* vol. 3, pp. 178-192, 2010.

[10] R. Eshuis and P. Grefen, "Structural Matching of BPEL Processes", *Proceedings of the Fifth European Conference on Web Services,* pp. 171-180, 2007.

[11] A. Wombacher and C. Li, "Alternative approaches for workflow similarity", *Proceedings of the 2010 IEEE International Conference on Services Computing,* pp. 337-345, 2010.

[12] R. Dijkman, M. Dumas, and L. García-Bañuelos, "Graph Matching Algorithms for Business Process Model Similarity Search", *Proceedings of the 7th International Conference on Business Process Management,* pp. 48-63, 2009.

[13] B. Yun, J. Yan, M. Liu, and Y. Yu, "Behavioral equivalence based web service discovery", *Proceedings of the 2008 International Conference on Computer Science and Software Engineering - volume 02,* pp. 368-371, 2008.

[14] S. Goedertiera, J. D. Weerdta, D. Martensa, J. Vanthienena, and B. Baesensa, "Process discovery in event logs: An application in the telecom industry", *Applied Soft Computing,* vol. 11, pp. 1697-1710, 2011.

[15] R. Nayak and B. Lee, "Web Service Discovery with additional Semantics and Clustering", in *Web Intelligence, IEEE/WIC/ACM International Conference on*, 2007, pp. 555-558.

[16] A. Ferro, R. Giugno, G. Pigola, A. Pulvirenti, D. Skripin, G. Bader, and D. Shasha, "NetMatch: a Cytoscape plugin for searching biological networks", *Bioinformatics,* pp. 910-912, 2007.

[17] R. Angell, G. Freund, and P. Willett, "Automatic spelling correction using a trigram similarity measure", *Information Processing & Management,* vol. 19, pp. 255 - 261, 1983.

[18] C. Figueroa, L. Sandino, and J. C. Corrales, "Plataforma para evaluar sistemas de recuperación de procesos de negocio", *Revista de Investigaciones UCM,* pp. 64-76, 2011.

[19] U. Küster and B. König-Ries, "Measures for Benchmarking Web Service Matchmaking Correctness", *Proceedings of the 7th international conference on The Semantic Web: research and Applications - Volume Part II,* pp. 45-59, 2010.

[20] P. Foggia, C. Sansone, and M. Vento, "A performance comparison of five algorithms for graph isomorphism", *in Proceedings of the 3rd IAPR TC-15 Workshop on Graph-based Representations in Pattern Recognition,* pp. 188-199, 2001.